

ユーザの既訪問スポットの位置づけに基づく未訪問スポットの説明手法

潘 健太[†] 北山 大輔[†]

[†] 工学院大学大学院工学研究科情報学専攻 〒163-8677 西新宿 1-24-2

E-mail: [†]tem18011@ns.kogakuin.ac.jp, ^{††}kitayama@cc.kogakuin.ac.jp

あらまし 近年、観光スポットを決める時に Web 上の観光情報を活用して計画を立てることが多くなっている。しかし、ユーザが多くエリアから訪問したいエリアを決めた上で、さらに自分のイメージに合う観光スポットを探すのは膨大な時間と労力を必要とする。また、ユーザが未訪問スポットに対して期待と不安を感じる場合がある。本研究では、ユーザの未知なスポットに対する理解を支援するためには、既に訪問したことがある観光スポットの特徴を未訪問スポットにあてはめて理解を支援する説明手法を提案する。観光スポット自身の特徴を重視するため、各観光スポットの特徴抽出に、ユーザが入力した観光スポットのすべてのレビュー、対象エリアの観光スポットのすべてのレビューを使用する。また、プロトタイプシステムを構築し、既訪問スポットと未訪問スポットとの説明情報の効果を検証する評価実験を行う。

キーワード 観光スポット, 理解支援, レビュー, 分散表現

1. はじめに

旅行先を決定する時、旅行者は観光スポット検索サイトや観光情報に関連する書籍を見て観光スポットを選び、旅行計画を立てる。しかし、ユーザにとって訪問したいエリアを決定した後、さらにエリア内に数多く存在する観光スポットから、自身のイメージから外れない観光スポットを見つけることは容易ではない。行きたい観光スポットが決まっていなかった場合はランキングやおすすめ情報を見て観光スポットを決めることが多くなると考えられる。この時、ユーザが選択した観光スポットに対するイメージが曖昧になるため不安を感じる場合がある。

近年、観光業とソーシャルネットワーキングサービスの発展スピードが加速しており、体験した観光スポットに対するレビューを観光スポット検索サイトに投稿しているユーザが増加している。さまざまな観光スポットを効果的に理解するためには、既存の情報をもとにして、未知な情報と既知な情報との対応関係を考えることが不可欠となる。この考え方は、以前に経験した事柄（ベースと呼ぶ）を、現在直面している事柄あるいは問題（ターゲットと呼ぶ）にあてはめる類推に相当する。たとえば、金沢の「にし茶屋街」という未知なスポットに対して既訪問の京都の「花見小路」と似ていると説明するとイメージの理解がしやすくなる可能性がある。

本研究では、ユーザの未訪問スポットに対する理解を支援するため、既に訪問したことがある観光スポットの特徴を未訪問スポットにあてはめて理解を支援する説明手法を提案する。具体的には、ユーザが入力した既訪問スポットと未訪問エリアから、レビューを用いて既訪問スポット内の各スポットの独特な特徴と未訪問エリア内の各スポットの独特な特徴を抽出し、比較を行って説明情報を提示する。このプロトタイプシステムにより、ユーザが未訪問スポットに対する理解の支援を目指す。図 1 は提案手法の概念図である。

本論文の構成は下記のとおりである。2 章では関連研究につ

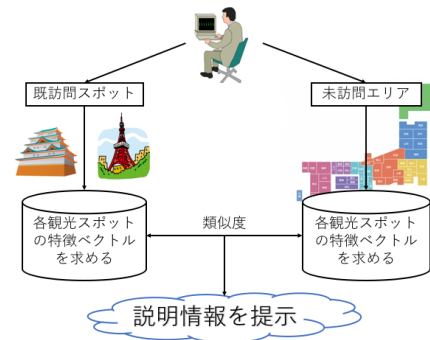


図 1 ユーザの既訪問スポットの位置づけに基づく未訪問スポット (エリア) の説明手法

いて述べる。3 章では提案手法の概要について述べる。4 章では構築したプロトタイプシステムの効果を検証する評価実験と考察について述べる。最後に 5 章ではまとめと今後の課題について述べる。

2. 関連研究

ユーザの体験履歴を利用した検索や推薦システムに関する研究が数多く発表されている。倉島ら [7] は、Flickr に投稿された写真のジオタグ情報を人々の旅行履歴として利用した旅行ルート推薦手法を提案した。この手法では、ユーザの現在地から行きやすい場所とユーザの興味に合致した場所に移動しやすいと仮定し、行動モデルを生成している。ユーザのジオタグ付き写真集合は、時間情報でソートすると個人の旅行履歴とみなすことができると考え、ジオタグ情報を利用してユーザの行動モデルを生成している。北村らは [8]、一般的な物体認識を用いて、過去の個人旅行写真から旅行計画のユーザの嗜好を推定することに基づいて観光地を推薦する方法を提案した。物体認識システムを用いて、写真で撮った被写体情報のキーワードを取得し、

グラフ視覚化技術によってキーワードの共起を表現した。また、グラフの視覚化技術に基づいて旅行写真付きのグラフを視覚化するユーザインターフェイスを紹介した。Cheng らは [9]、自由に利用できるコミュニティ寄稿の写真を活用して、パーソナライズされた旅行のおすすめ焦点を当て、特定のユーザープロフィールまたは属性を考慮し、パーソナライズされた旅行の推奨を行うことを提案した。

類推は創造的思考に貢献すると指摘されてた [1]。既知の知識(ベースと呼ぶ)から概念(ターゲットと呼ぶ)を獲得するときに類推思考が働くと考えられる [2]。類推に関する研究の多くは、ベースとなる学習データとターゲットとなる問題が与えられ、物事の特徴を問題の特徴にマッピングして問題を解決するもの [3] である。Gick らは、不確定な問題の解を見つけるためのガイドとして、異種ドメイン間の類推の使用を調査するように設計した。学習データの与え方や機能について研究したもの [4] や、認知的な熟達度に応じて問題を解決するかどうかを明らかにしたもの [5] がある。これらを含む従来研究の多くにおいては、類推に用いるベースとターゲットを与えたうえで、一定の手順に従って問題解決を行っている。また、構造の類似性には 3 種類あり、特徴の共有数で決まる「対象レベルの類似性」、ベースに存在する関係とターゲットに存在する関係の共有度に基づく「関係レベルの類似性」、および題の解法あるいは目標レベルでの類似性である「プラグマティックな類似性」とがある。

従来のユーザの体験履歴を利用する手法では、履歴写真のジオタグ情報を分析し、ユーザの嗜好とする研究が多く行われている。また、類推技術に関して学習支援でよく使われている。本研究では、既訪問スポットと未訪問スポットのレビューを使って、ユーザ既訪問スポット集合と未訪問スポット集合のそれぞれ集合の各スポットの相対的な特徴を求め、関連付けることによって、スポットに対する理解を支援するため説明情報を提示することができる。また、本研究では、類推の質を明示的扱うため、構造の類似性「関係レベルの類似性」に近いと考えられる。

3. 未訪問スポットの説明手法

我々は、ユーザの既訪問スポットの位置づけに基づく未訪問スポットの説明手法を提案する。具体的にはまず、ユーザが既訪問の複数個の観光スポットと訪問したい観光スポットエリア情報を入力する。既訪問スポットレビューベクトルを使って既訪問スポット毎の特徴ベクトルを求める。未訪問スポットも同様にエリア内の各スポットの特徴ベクトルを求める。次に、既訪問スポットレビューベクトルと未訪問スポットレビューベクトルの差分特徴に類似する特徴を持つ未訪問観光スポット関連付けを行う。最後に、TFIDF を用いて未訪問スポットの理解支援のための説明手法を定義し、ユーザに提示する。

3.1 スポットのレビューから特徴ベクトル生成

既訪問スポットや未訪問スポットのレビューベクトルは、形態素解析器「mecab-ipadic-NEologd」^(注1)で分かち書き(原型)

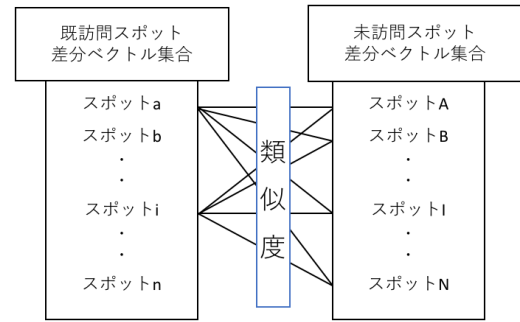


図2 類似度計算概念図

したレビューを利用して作成する。その後、Doc2Vec^(注2)のDistributed Bag-of-Words を利用して、各スポットの全レビューを使って 300 次元で作成したベクトルを使う。本稿に置いて、レビューデータは 2016 年 09 月末までじゃらん^(注3)から取得したものをを用いる。

3.2 観光スポットの役割に関する相対的な特徴

本研究では、観光スポットの特徴は相対的な特徴を利用する。相対的な特徴とは、特定の観光スポットが、ある観光スポット集合に含まれた他の観光スポットと比較した場合における独特な特徴である。例として、観光スポット集合内に鹿苑寺と清水寺が存在する場合を考える。このとき鹿苑寺の特徴は、金色、金箔、輝き等となり、清水寺の特徴は、舞台、胎内、一望等となる。どちらも京都に存在する寺院であるため、京都や寺院に関連する特徴は独特な特徴として現れることがない。次に、観光スポット集合内に東京都庁舎展望台と鹿苑寺が存在する場合を考える。このとき鹿苑寺の特徴は、金閣寺、お寺、金色、京都等となり、東京都庁舎の特徴は、展望、夜景、新宿等となる。観光スポットのカテゴリが大きく異なる場合であれば、カテゴリとしての特徴が現れる。また、スポット自身の特徴を表すことができる。本研究では、あるスポットが集合内の他のスポットと比較するとき、より各スポットの特徴を明らかにできる相対的な特徴に着目して研究を行う。

スポット差分ベクトルは式 1 として定義される。スポット差分ベクトルを求めるスポットを除いたスポット集合の各スポットのスポットベクトルの平均値を引いた値となる。 $spot_{set} = \{s_1, s_2, \dots, s_i, \dots, s_n\}$ は既訪問スポット集合や未訪問スポット集合となっている。また、 s_i は集合内のある観光スポットを示している。

$$v_i = s_i - average(spot_{set} - s_i) \quad (1)$$

3.3 説明スポットの決定

既訪問スポットの各特徴差分ベクトル v_i と未訪問スポットの各特徴差分ベクトル v_j から、既訪問スポットと未訪問スポット間の相対的な特徴の類似度(図2)を求める。類似度計算に

(注2) : <https://radimrehurek.com/gensim/models/doc2vec.html>

(注3) : <https://www.jalan.net/kankou/>

(注1) : <https://github.com/neologd/mecab-ipadic-neologd/>

表 1 形態素解析の例

レビュー文書	園内も広く、気分転換に散歩したりするのにちょうどよい。きれいに清掃などもされていて、気分がよいです。
形態素解析	園内 広い 気分転換 散歩 ちょうど よい きれい 清掃 きれい 気分 よい

は、コサイン尺度 (式 2) を用いる。

$$\cos(v_i, v_j) = \frac{v_{i1}v_{j1} + v_{i2}v_{j2} + \dots + v_{in}v_{jn}}{\sqrt{v_{i1}^2 + \dots + v_{in}^2} \times \sqrt{v_{j1}^2 + \dots + v_{jn}^2}} \quad (2)$$

既訪問スポットの各特徴ベクトルと未訪問エリア内の各特徴ベクトルの類似度が 0.125 以上かつ、類似度が最も高い既訪問スポットと未訪問スポットの関連付けを行う。また、既訪問スポットと未訪問スポットとの関連をつけるとき以下の 2 つの方法がある。

(1) 既訪問スポットをベースにして未訪問スポットと関連付ける方法

(2) 未訪問スポットをベースにして既訪問スポットと関連付ける方法

方法 1 では、既訪問スポットをベースにすると未訪問エリア内のスポットが複数の特徴を保持する場合は同じ未訪問スポットと関連づける場合がある。方法 2 では、未訪問スポットをベースにすることによって、既訪問スポット集合から各スポットの特徴を取り出し、未訪問スポットと関連付けることができるため、方法 2 を利用する。また、本研究では、未訪問スポットを説明するため未訪問スポットをベースにすることが妥当と考えられる。

3.4 説明するための役割語の抽出

観光スポットのレビューはすべて形態素解析器「mecab-ipadic-NEologd」を使用することで、単語抽出処理を行う。しかし、これらを用いて得られた単語は、日本語として成立しない語が含まれており、これらノイズの削除が必要となる。具体的には、助詞、助動詞、連体詞、記号、ストップワードを削除する (表 1)。

節 3.3 で関連付けした既訪問スポットと未訪問スポットの情報は単語形式でユーザに提示するため、ある観光スポットのレビュー集合を文書 i とし、 i に対する単語 j が出現するスポット集合の出現回数を $TF_{i,j}$ 、単語 j がスポット集合の文書数を DF_j 、スポット集合内の全スポット数を $|D|$ とした時、そのスポットにおける単語の特徴量は、式 3 で定義される。

$$word_{i,j} = TF_{i,j} \times IDF_j \quad (3)$$

$$IDF_j = \log\left(\frac{|D|}{DF_j}\right) \quad (4)$$

本手法では、既訪問スポットに関して、ユーザが複数個のスポットを入力する。それぞれのスポットの全レビューをまとめて 1 つの文書と見なし、それ以外のスポットの全レビューも文書とみなすことで、式 3、4 によって TFIDF 値を算出し、既訪問スポット毎の特徴語とする。

未訪問エリアに関して、ユーザがエリアを指定して入力する。



図 3 説明手法インターフェース

エリア内のそれぞれのスポットの全レビューをまとめて 1 つの文書と見なし、それ以外のスポットの全レビューも文書とみなすことで、式 3、4 によって TFIDF 値を算出し、未訪問エリアのスポット毎の特徴語とする。

関連付けした既訪問スポットと未訪問スポットの説明情報は、TFIDF で求めた各スポットの特徴語による調和平均を用いて決定する。調和平均とは、逆数の算術平均の逆数である。既訪問スポットのレビュー文書と、未訪問スポットのレビュー文書に、共通して出現する単語を抽出する。抽出した単語のスコアは式 5 によって定義する。 $word_{familiar}$ と $word_{unfamiliar}$ は同じ単語がそれぞれ既訪問スポットの TFIDF 値と未訪問スポットの TFIDF 値を示している。単語スコアの値が大きくと既訪問スポットと未訪問スポットのそれぞれの TFIDF 値が大きい、つまり単語がそれぞれの文書に置いて重要度が高いことを示している。よって、単語スコアの上位 10 個の単語を説明情報としてユーザに提示する (図 3)。

$$score = \frac{1}{\frac{1}{2} \left(\frac{1}{word_{familiar}} + \frac{1}{word_{unfamiliar}} \right)} \quad (5)$$

3.5 未訪問スポットの説明情報の例

表 2 は、ユーザが既に訪問したことがあるスポットと、未訪問スポットの集合の例である。未訪問スポットは東京都内からランダムに選んだ 5 つのスポットである。節 3. の提案手法を使って説明単語を求めた結果は表 3 である。

公園という特徴に注目すると、未訪問スポット集合内で最も公園に近いスポットは新宿御苑であると考えられる。既訪問スポット集合内には小田原城址公園と奈良公園の二つの公園がある。小田原城址公園には花や遊具に対する記述が多く、奈良公園は鹿や草に対する記述が多い。新宿御苑は花や遊具に対する記述が多いため、小田原城址公園との関連性があると考えられる。

三島スカイウォークは既訪問スポット集合内で眺めがいい、高い特徴が 1 番強いことが考えられる。未訪問スポット集合内で眺めがいい、高い特徴が強いのは東京スカイツリーや東京タワー大展望台の 2 つがある。東京スカイツリーはより高いことによって三島スカイウォークと考えられるが、説明情報を見ると富士山が入っている。2 つスポットの高さは富士山という単語によって強調されている。提案手法は各集合のスポット毎の特徴を表すことができる。

表 2 既訪問スポット集合と未訪問スポット集合

既訪問スポット名	未訪問スポット名
浅草寺	東京ディズニーランド (R)
小田原城址公園	新宿御苑
伏見稲荷大社	東京スカイツリー
奈良公園	東京タワー大展望台
三島スカイウォーク	明治神宮

4. 評価実験

4.1 実験内容

絶対的な特徴と提案手法の相対的な特徴による説明情報を提示する方法の比較を行う。また、相対的な特徴を利用する場合において、調和平均を利用して単語スコアを算出する方法と相加平均を利用して単語スコアを算出する方法の比較を行った。

クラウドソーシングのサービスである、CrowdWorks^(注4)を利用して 24 人の被験者を集めた。じゃらんで取得した観光スポットを使って、被験者の既訪問スポットの選択による未訪問スポットの説明情報の提示を行った。絶対的な特徴と相対的な特徴の説明情報のパターンは以下の 4 つとなる。

- 絶対的な特徴 (カテゴリ, 滞在時間, 訪問時期)
- 絶対的な特徴 (特徴ベクトル)
- 相対的な特徴 (差分ベクトル, 調和平均を利用して単語スコアを算出)
- 相対的な特徴 (差分ベクトル, 相加平均を利用して単語スコアを算出)

説明情報 A は、観光スポット検索サイトでスポットを検索するため使う絞り込み情報である。絞り込み情報は例は以下の 3 つである。

- ・ カテゴリ：神社・神宮・寺院, 観光施設・名所巡り等
- ・ 滞在時間：1 時間未満, 1~2 時間等
- ・ 訪問時期：1~12 月, 春, 夏, 秋, 冬

まず、既訪問スポットを使って、未訪問スポットとのカテゴリが一致かどうか、滞在時間と一致かどうか、訪問時期と一致かどうかの順で徐々に情報を絞って、既訪問スポットと未訪問スポットの対応付けを行う。次に、絞った後の未訪問スポットが複数残った場合、レビュー数が 1 番多い未訪問スポットを利用する。最後に、節 3.4 を使って説明するための役割語を抽出し、被験者に提示する。説明情報 B は、節 3.1 で作成した特徴ベクトルを使って、各スポットの特徴とする。説明情報 C は、提案手法である。説明情報 D は、関連付けした既訪問スポットと未訪問スポットの説明情報は、TFIDF で求めた各スポットの特徴語による調和平均を用いて決定する。既訪問スポットのレビュー文書と、未訪問スポットのレビュー文書に、共通して出現する単語を抽出する。また、抽出した単語のはそれぞれのスポットの TFIDF の値の平均を以上である必要がある。既訪問スポット単語の TFIDF 値と未訪問スポット単語の TFIDF 値の差の絶対値を単語スコアとして算出し、単語スコアが最も 0 に近い 10 個の単語を説明情報として被験者に提示する。

まず、被験者は既に訪問したことがあり、気に入った観光スポットを 4 つ以上 10 つ以下入力した。入力するスポットは、検索候補から選んでもらう。例えば、新宿御苑—新宿, 清水寺—京都などがある。次に、被験者は旅行等で行ったことがなく、これから行ってみたい都道府県・エリアを入力した。説明情報のパターンの A から D の順で処理を行って、ユーザに未訪問スポット名、キーワード、既訪問スポット名を提示し、以下の 5 つの評価から 1 つを選択した。

- (1) キーワードなし
- (2) 2 つのスポットにそもそも関係性があるが、キーワードにより関係が明確になった
- (3) 2 つのスポットの関係にキーワードにより初めて気がついた
- (4) 2 つのスポットに関係性があるがキーワードは関係性を表していない
- (5) 2 つのスポットに関係性はない

4.2 実験結果と考察

表 4 は評価 1 から 5 の説明情報 A から D のそれぞれの実験結果のデータの数である。今回の実験では、使用可能のデータの合計は 394 件である。説明情報 A の方が未訪問スポットと関連がある既訪問スポットの数が最も少ない。説明情報 B の方が未訪問スポットと関連がある既訪問スポットの数が最も多い。説明情報 C と説明情報 D について、未訪問スポットと既訪問スポットの関連付け手法は同じであるため、数が同じとなる。

表 5 は説明情報 A から D において評価 1 から 5 の実験結果のデータの数の割合である。被験者に提示する説明情報 A は、未訪問スポットと既訪問スポットがそもそも関連性があり、またキーワードを提示することによってさらに明確となった。説明情報 D は、未訪問スポットと既訪問スポットに関連性がないと思っていたが、被験者にキーワードを提示することによって初めて気がついたため、本研究の提案手法に関連性がある。説明情報 C と説明情報 D は、被験者に提示する未訪問スポットと既訪問スポットがそもそも関連性がない割合が多い。4 つの説明情報パターンにおいて説明情報 D のみ被験者に提示するキーワードがない。よって、調和平均を利用することによって被験者にキーワードを提示することができる場合が多いといえる。また、評価 1, 評価 4 と評価 5 は説明情報を意味がなしていないことが示すことができる。よって、説明情報 D はもっとの意味をなしていないことがいえる。

評価 1 と評価 2 はから、被験者に提示する未訪問スポットと既訪問スポットがそもそも関連性があるカテゴリを利用した場合が 1 番良い。しかし、関連性のあるスポットの数はカテゴリに制限しているため被験者に提示できる数も少ない。また、別カテゴリのスポット情報を被験者に提示できないため意外性を求めることができないといえる。2 番目は特徴ベクトル利用する場合、3 番目は差分ベクトルと調和平均の組み合わせるとき、最も良くないのは説明情報 D の差分ベクトルと相加平均の組み合わせである。説明情報 C と説明情報 D の関連性があるスポットの数が同じであるが、相加平均を使って単語スコアを算出することによって被験者に提示するキーワードの理解支援

(注4) : <https://crowdworks.jp/>

表 3 説明情報の例

未訪問スポット	既訪問スポット	説明情報
新宿御苑	小田原城址公園	お花見, 咲き誇る, 園内, 桜の時, のんびり, 手入れ, 自然, 遊具, ツツジ
東京スカイツリー	三島スカイウォーク	富士山, 揺れ, 高所恐怖症, 揺れる, 天井, 絶景, エレベーター, パノラマ, 展望デッキ, 昇る

表 4 実験結果のデータ数の統計

評価	説明情報 A	説明情報 B	説明情報 C	説明情報 D	合計
1	0	0	0	4	4
2	19	44	32	26	121
3	20	62	53	56	191
4	1	3	3	3	10
5	6	21	21	20	68
合計	46	130	109	109	394

表 5 説明情報における評価の割合

評価	説明情報 A	説明情報 B	説明情報 C	説明情報 D
1	0.00%	0.00%	0.00%	3.67%
2	41.30%	33.85%	29.36%	23.85%
3	43.48%	47.69%	48.62%	51.38%
4	2.17%	2.31%	2.75%	2.75%
5	13.04%	16.15%	19.27%	18.35%

の役割が減少したといえる。

本研究の提案手法を示している, 評価 2 は差分ベクトルと相加平均の組み合わせが最も良い, 差分ベクトルと調和平均の組み合わせが 2 番目良い, 3 番目は特徴ベクトル利用する場合, 最も良くないのはカテゴリを用いる場合である。よって, 説明情報 B の特徴ベクトルと説明情報 C の差分ベクトルと調和平均の組み合わせが良い結果を示すことができるといえる。

カテゴリは

5. まとめと今後の課題

本研究では, ユーザが行きたい観光スポットが決まっていな場合, ランキング, おすすめ情報やカテゴリなどに観光検索情報を使用することによって, 検索した観光スポットがに対する理解が困難であることを着目した。未訪問なスポットに対する理解を支援するために, ユーザが既に訪問したことがあるスポットの特徴を未訪問スポットにあてはめて理解を支援する説明手法を提案した。

評価実験では 4 つの説明情報パターンを用いて比較を行った。結果, カテゴリを利用する場合では, 未訪問スポットと関連する既訪問スポットが最も少ない。差分ベクトルと相加平均を利用する場合では, キーワードの理解を支援する役割が最も少ない。

謝 辞

文 献

- [1] K. J. Holyoak and P. Thagard, “Mental Leaps: Analogy in Creative Thought, MIT Press”, Journal of Japanese Society for Artificial Intelligence, Vol.11, No.3, pp.489, 1996
- [2] D. Gentner, “Structure-Mapping: A Theoretical Framework

- for Analogy”, Cognitive Science, Vol.7, pp.155–170, 1983
- [3] M. L. Gick and K. J. Holyoak, “Analogical Problem Solving”, Cognitive Psychology, Vol.12, pp.306–355, 1980
- [4] M. L. Gick and K. J. Holyoak, “Scheme Induction and Similarity in Analogical Transfer”, Cognitive Psychology, Vol.15, pp.1–38, 1983
- [5] Z. Chen and M. W. Daehler, “Positive and Negative Transfer in Analogical Problem-solving by 6-years-old Children”, Cognitive Development, Vol.4, No.4, pp.327–344, 1989
- [6] K. J. Holyoak and P. Thagard, “Analogical Mapping by Constraint Satisfaction”, Cognitive Science, Vol.13, pp.295–355, 1989
- [7] T. Kurashima, T. Iwata, G. Irie and K. Fujimura., “Travel route recommendation using geotags in photo sharing sites”, CIKM '10 Proceedings of the 19th ACM international conference on Information and knowledge management, pp.579–588, 2010
- [8] R. Kitamura and T. Itoh, “Tourist Spot Recommendation Applying Generic Object Recognition with Travel Photos”, ITE Tech. Rep., Vol.42, No.12, AIT2018-94, pp.185–188, 2018
- [9] A. J. Cheng, Y. Y. Chen, Y. T. Huang and Winston H. Hsu, “Personalized Travel Recommendation by Mining People Attributes from Community-Contributed Photos”, MM '11 Proceedings of the 19th ACM international conference on Multimedia, pp.83–92, 2011
- [10] T. Kudo, K. Yamamoto and Y. Matsumoto, “Applying Conditional Random Fields to Japanese Morphological Analysis”, Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004), pp.230–237, 2004